Spatially Balanced Sampling using the Halton Sequence

Blair Robertson* Jennifer Brown† Trent McDonald‡ Chris Price§

Abstract

A spatial sampling design determines where sample locations are placed in a study area. The main objective is to select sample locations in such a way that valid scientific inferences can be made to all regions of the study area. A sample that is well-spread over the study area is called a spatially balanced sample. Spatially balanced sampling designs are known to be efficient when surveying natural resources because nearby locations tend to be similar. This paper shows how the Halton sequence can be used to draw spatially balanced samples from environmental resources.

Key Words: Balanced acceptance sampling (BAS), environmental sampling, Halton iterative partitioning (HIP), over-sampling, SDraw

1. Introduction

A spatial sampling design determines where sample locations are placed in a study area. The main objective of a spatial design is to draw sample locations in such a way that valid scientific inferences can be made to all regions of the study area (McDonald 2014). To achieve good estimates of population characteristics, the spatial pattern of the sample should be similar to the spatial pattern of the population. However, the spatial pattern of the response variable is usually not known. Fortunately, when sampling natural resources, nearby locations tend to be similar because they interact with one another and are influenced by the same set of factors (Stevens & Olsen 2004). This means sample efficiency can be increased by spreading sample locations evenly over the resource. Stevens and Olsen (2004) called well spread samples *spatially balanced samples* and measured spatial balance using the Voronoi tessellation of a sample.

Consider drawing n sample locations from an continuous resource $\Omega \subset [0,1)^2$ with $\lambda(\Omega) > 0$, where λ is the Lebesgue measure. Let $\pi(\boldsymbol{x}) = nf(\boldsymbol{x})$ be an inclusion density function, where $f(\boldsymbol{x}) : [0,1)^2 \to \mathbb{R}_{\geqslant 0}$ is a bounded probability density function such that

$$\int_{\Omega} f(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

A sample, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset \Omega$, is considered spatially balanced if

$$v_i = \int_{\omega_i} \pi(\boldsymbol{x}) d\boldsymbol{x} \approx 1 \quad \text{for all } i = 1, 2, \ldots, n,$$

where ω_i is the Voronoi polygon for $\boldsymbol{x}_i$

$$\omega_i = \big\{\boldsymbol{x} \in [0,1)^2 : \|\boldsymbol{x} - \boldsymbol{x}_i\| \leqslant \|\boldsymbol{x} - \boldsymbol{x}_j\| \text{ for all } j = 1, 2, \ldots, n\big\}.$$

The spatial balance of $n = 50$ points drawn from $\Omega = [0,1)^2$ is illustrated in Figure 1.

The spatial balance of a sample drawn from a discrete population is measured in a similar way. Let U be a finite population of N points from $[0,1)^2$ and let $0 < \pi_i < 1$ denote

*School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
†School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
‡Western EcoSystems Technology, Wyoming, USA
§School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

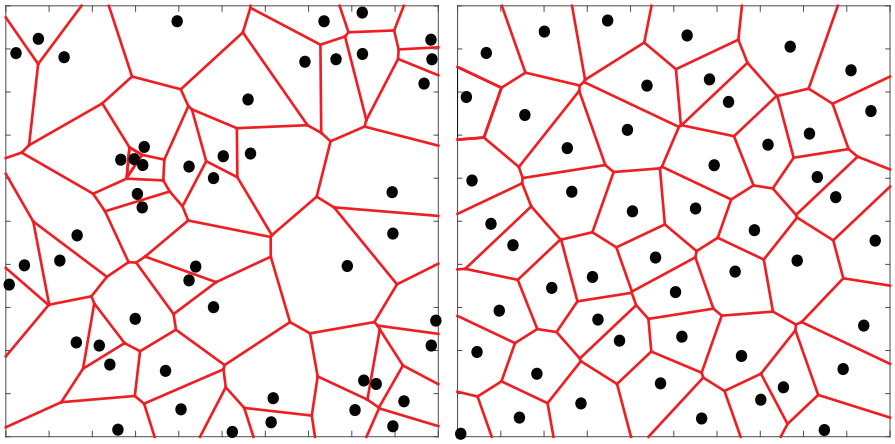

Figure 1: (Left) a simple random sample (SRS) of $n = 50$ points drawn from $\Omega = [0, 1)^2$. (Right) an equal probability BAS sample of $n = 50$ points drawn from $\Omega = [0, 1)^2$. In this case, v_i is proportional to the area of ω_i (shown in red). BAS has far better spatial balance than SRS because the areas of each ω_i are more similar in size.

the inclusion probability of $\boldsymbol{x}_i$ such that $\sum_{i=1}^{N} \pi_i = n$. A sample, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subset U$, is considered spatially balanced if

$$v_i = \sum_{\boldsymbol{x}_j \in \omega_i} \pi_j \approx 1 \quad \text{for all } i = 1, 2, \ldots, n,$$

where ω_i is the Voronoi set for $\boldsymbol{x}_i$

$$\omega_i = \left\{\boldsymbol{x} \in U : \|\boldsymbol{x} - \boldsymbol{x}_i\| \leqslant \|\boldsymbol{x} - \boldsymbol{x}_j\| \text{ for all } j = 1, 2, \ldots, n\right\}.$$

Stevens and Olsen (2004) introduced the phrase *spatially balanced sampling* and presented the first spatially balanced design. Their design, called Generalized Random Tessellation Stratified (GRTS), is a popular sampling design and it is frequently used in environmental monitoring (Kermorvant et al. 2019). GRTS can be applied to point, linear and continuous resources and can draw unequal probability samples. Another useful property of GRTS is its ability to dynamically add points from a spatially balanced over-sample to the sample as non-target or inaccessible points are discovered (Stevens & Olsen 2004; Larsen, Olsen & Stevens 2008). This feature is popular with field researchers because the largest sample that their budget permits can be analysed, but it does not eliminate issues associated with non-response (Robertson et al. 2018).

Spatially balanced sampling has been an active area research over the past decade and a variety of designs have been proposed, where each design uses a different strategy to achieve spatial balance. The Local Pivotal Method (LPM; Grafström, Lundström & Schelin 2012), is an application of the Pivotal Method (Deville & Tillé 1998) that gives spatially balanced samples. Grafström (2012) also modified correlated Poisson sampling (Bondesson & Thorburn 2008) to draw spatially balanced samples, called spatially correlated Poisson sampling (SCPS). LPM is algorithmically simpler than SCPS, but SCPS can achieve a higher degree of spatial balance (Grafström & Schelin 2014). Benedetti & Piersimoni (2017) presented a flexible class of spatially balanced designs that draw their samples based on a within-sample distance. Balanced Acceptance Sampling (BAS; Robertson et al. 2013), its modified version (Robertson et al. 2017) and Halton Iterative Partitioning

(HIP; Robertson et al. 2018) all use the Halton sequence (Halton 1960) to draw spatially balanced samples. This paper focuses on BAS and HIP, and how these designs use the Halton sequence to draw spatially balanced samples.

The remainder of this paper is organised as follows. Section 2 introduces the Halton sequence and describes properties of the sequence that are pertinent to BAS and HIP. The sampling designs BAS and HIP are described in Section 3 and concluding remarks are given in Section 4

2. Halton Sequence

The Halton sequence is a quasi-random number sequence which spreads points evenly over the unit box in relatively low dimensions ($d \leqslant 10$). Quasi-random sequences have been used as a substitute for random numbers in many fields, including numerical integration (Niederreiter 1978, 2003), numerical optimization (Sobol 1979; Torn & Žilinskas 1989; Price & Price, 2012; Robertson, Price & Reale, 2014) and environmental sampling (Robertson et al. 2013, 2017, 2018). The Halton sequence is particularly useful in these fields because it is simple, generates evenly spread points (see Figure 2) and has similar spatial properties to a regular grid or lattice. However, unlike a regular lattice, points can be added incrementally with no clumping of points.

The random-start Halton sequence $\{\boldsymbol{x}_j\}_{j=1}^{\infty}$ in $[0,1)^d$ is defined as follows. The ith coordinate of the jth point in this sequence is (Price & Price 2012)

$$x_j^{(i)} = \sum_{p=0}^{\infty} \left\{ \left\lfloor \frac{u_i + j}{b_i^p} \right\rfloor \bmod b_i \right\} \frac{1}{b_i^{p+1}},$$

where u_i is a random non-negative integer, b_i is a positive integer and $\lfloor \cdot \rfloor$ is the floor function. The bases b_i are chosen to be small, co-prime integers to ensure the points are evenly spread over the unit box. In this paper, the bases are the first d prime numbers. The random-start Halton sequence is

$$\{\boldsymbol{x}_j\}_{j=1}^{\infty} = \left\{ (x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(d)}) \right\}_{j=1}^{\infty}. \tag{1}$$

Let $B = \prod_{i=1}^{d} b_i^{J_i}$, where J_i is any non-negative integer. It can be shown (Price & Price 2012; Robertson et al. 2017, 2018) that B consecutive points from (1) will have exactly one point in each of the Halton boxes defined by

$$\bigtimes_{i=1}^{d} \left[m_i b_i^{-J_i}, (m_i + 1) b_i^{-J_i} \right), \tag{2}$$

where m_i is an integer satisfying $0 \leqslant m_i < b_i^{J_i}$, for all $i = 1, \ldots, d$ (see Figure 2). The sequence is also quasi-periodic (with period B) because points of the form $\boldsymbol{x}_{j+\alpha B}$ with $\alpha = 0, 1, \ldots$, are in the same box (Robertson et al. 2017, 2018). Hence, contiguous subsequences from (1) are also spatially balanced (see Figure 2).

3. BAS and HIP

Balanced acceptance sampling (BAS; Robertson et al. 2013, 2017) is a spatially balanced sampling design that draws its sample using the random-start Halton sequence (1). Consider drawing n sample locations from an continuous resource $\Omega \subset [0,1]^2$ with $\lambda(\Omega) > 0$, where λ is the Lebesgue measure. An equal probability BAS sample is simply the first n

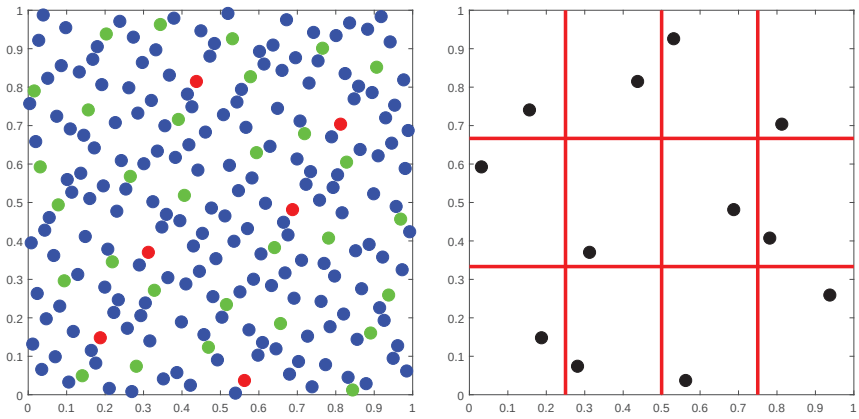

Figure 2: (Left) A random-start Halton sequence of $n = 216$ points with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_6$ in red and $\boldsymbol{x}_7, \ldots, \boldsymbol{x}_{36}$ in green. This colouring illustrates how contiguous sub-sequences from (1) are also spatially balanced. (Right) Halton boxes (2) for $B = 2^2 \times 3 = 12$ and B consecutive points from (1), where each box has exactly one point.

points from (1) that fall within Ω. However, if $\boldsymbol{x}_1 \notin \Omega$, discard the sequence and generate another (Robertson et al. 2017). Unequal inclusion probability samples can also be drawn using a rejection sampling technique (Robertson et al. 2013). BAS is conceptually simple, computationally efficient, can be used to draw spatially balanced over-samples/master samples (van Dam-Bates, Gansell & Robertson 2018) and performs well on continuous resources (Robertson et al. 2018).

BAS can also be applied to point and linear resources. However, there are two potential drawbacks when sampling these resources with BAS. First, acceptance/rejection sampling is required to draw its sample, and hence targeted inclusion probabilities are not necessarily achieved. Robertson et al. (2017) provided a simple modification to BAS that achieved targeted inclusion probabilities in specific settings and reduced discrepancies in the general setting. Despite this potential drawback, the actual inclusion probabilities can be computed for unbiased estimation.

The second potential drawback of sampling point resources with BAS is the sampling frame that BAS uses. BAS constructs its frame by replacing the N points in $[0, 1)^2$ with N non-overlapping boxes of equal size, with one point in each box. The BAS sample of size n is then drawn as follows. First, a random-start Halton sequence is defined

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\} \subset [0, 1)^2, \tag{3}$$

where k is chosen so that n boxes contain at least one point from (3). The points from the resource within these n boxes define the BAS sample. However, if the point resource is large or lacks grid structure, BAS can become inefficient because the boxes tend to be small and k can be enormous (Robertson et al. 2018).

Halton iterative partitioning (HIP; Robertson et al. 2018) extends BAS to better handle point resources. HIP iteratively partitions a resource into $B \geqslant n$ nested boxes using the quasi-periodic property of the Halton sequence. These boxes have the same nested structure as (2), but different sizes (see Figure 3). These boxes are then uniquely numbered using a random-start Halton sequence of length B. The HIP sample is obtained by randomly drawing one point from each of the boxes numbered $1, 2, \ldots, n$. This procedure is illustrated in Figure 3. Unequal probability samples can be drawn by altering the inclusion probability of each box (Robertson et al. 2018).

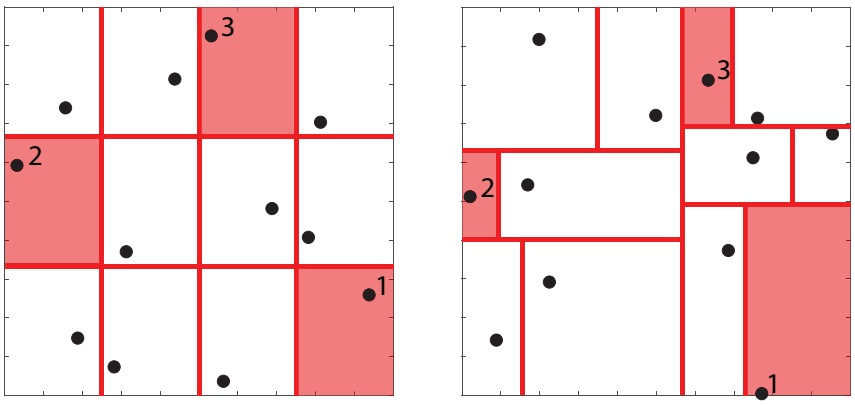

Figure 3: (Left) Halton boxes for $B = 2^2 \times 3 = 12$ and B consecutive points from (1). The first three points from the sequence are numbered. (Right) A HIP partition for a $N = 12$ point resource using $B = 12$ boxes. This partition has the same nested structure as the Halton boxes, but the boxes have different sizes. Using the box numbering from the left figure, an equal probability HIP sample of $n = 3$ points from the resource is shown.

The HIP design is conceptually simple, computationally efficient on large N point resources, has a rapid implementation for equal probability sampling and is embarrassingly parallel. It can be applied to continuous and point resources and achieves targeted inclusion probabilities/density. HIP samples use the same ordering as the Halton sequence to ensure contiguous sub-samples are spatially balanced. This feature makes HIP particularly useful for spatially balanced over-sampling if non-target or inaccessible units are discovered.

BAS and HIP samples can be drawn using the SDraw package (McDonald 2016) in the R programming language. This package allows spatially balanced samples/over-samples to be drawn from point, linear and continuous resources.

4. Conclusion

Spatially balanced sampling designs are commonly used for sampling natural resources and a variety of designs have been proposed. BAS and HIP are spatially balanced designs that use the Halton sequence to draw their samples. BAS uses points from the sequence to draw its sample and HIP uses properties of the sequence to partition the resource before the sample is drawn. The potential advantages of these designs over other spatially balanced designs include being conceptually simple, computationally efficient and being able to draw spatially balanced over-samples. This makes them particularly useful for sampling natural resources because imperfect sampling frames and accessibility problems result in fewer units being observed than planned. Although spatially balanced over-sampling achieves the desired sample size and is popular with field researchers, it will not eliminate the non-response or the bias of an inference. BAS and HIP samples can be drawn using the SDraw package in the R programming language.

REFERENCES

Benedetti, R. and Piersimoni, F. (2017). "A spatially balanced design with probability function proportional to the within sample distance," *Biometrical Journal*, 59, 1067–1084.

Bondesson, L. and Thorburn, D. (2008). "A list sequential sampling method suitable for real-time sampling," *Scandinavian Journal of Statistics*, 35, 466–483.

Deville, J. C. and Till, Y. (1998). "Unequal probability sampling without replacement through a splitting method," *Biometrika*, 85, 89–101.

Grafström, A., Lundström, N.L.P. and Schelin, L. (2012). "Spatially balanced sampling through the pivotal method," *Biometrics*, 68, 514–520.

Grafström, A. (2012). "Spatially correlated Poisson sampling," *Journal of Statistical Planning and Inference*, 142, 139–147.

Grafström, A. and Schelin, L. (2014). "How to select representative samples," *Scandinavian Journal of Statistics*, 41, 277–290.

Halton, J. H. (1960). "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals," *Numerische Mathematik*, 2, 84–90.

Kermorvant, C., D'Amico, F., Robertson, B., Bru, N. and Caill-Milly, N. (2019). "Spatially balanced sampling designs for environmental surveys," *Environmental Monitoring and Assessment*, 191: 524. https://doi.org/10.1007/s10661-019-7666-y

Larsen, D. P., Olsen, A. R. and Jr. Stevens, D. L. (2008). "Using a master sample to integrate stream monitoring programs," *Journal of Agricultural, Biological and Environmental Statistics*, 13, 243–254.

McDonald, T. (2014). "Sampling Designs for Environmental Monitoring," In B.F.J. Manly, J.A. Navarro Alberto (Ed.), Introduction to Ecological Sampling, Florida, CRC Press, Taylor and Francis Group.

McDonald, T. (2016). "SDraw: spatially balanced sample draws for spatial objects," R package version 2.1.8. https://CRAN.Rproject.org/package=SDraw

Niederreiter, H. (1978). "Quasi-monte carlo methods and pseudo-random numbers," *Bulletin of the American Mathematical Society*, 84, 957–1041.

Niederreiter, H. (2003). "Error bounds for quasi-monte carlo integration with uniform point sets," *Journal of Computational and Applied Mathematics*, 150, 283–292.

Price, C. J. and Price, C. P. (2012). "Recycling primes in the Halton sequence: an optimization perspective," *Advanced Modelling and Optimization*, 14, 17–29.

Robertson, B. L., Brown, J. A., McDonald, T. and Jaksons, P. (2013). "BAS: Balanced acceptance sampling of natural resources," *Biometrics*, 3, 776–784.

Robertson, B. L., Price, C. J. and Reale, M. (2014). "A CARTopt method for bound constrained global optimization," *ANZIAM Journal*, 55, 109–128.

Robertson, B. L., McDonald, T., Price, C. J. and Brown, J. A. (2017). "A modification of balanced acceptance sampling," *Statistics and Probability Letters*, 129, 107–112.

Robertson, B. L., McDonald, T., Price, C. J. and Brown, J. A. (2018). "Halton iterative partitioning: spatially balanced sampling via partitioning," *Environmental and Ecological Statistics*, 3, 305–323.

Sobol, I. M. (1979). "On the systematic search in a hypercube," *SIAM Journal of Numerical Analysis*, 16, 790–793.

Stevens, D. L., Jr. and Olsen, A. R. (2004). "Spatially balanced sampling of natural resources," *Journal of the American Statistical Association*, 99, 262–278.

Torn, A. and Žilinskas, A. (1989). "Global Optimization, Lecture Notes in Computer Science," Vol. 350. Springer-Verlag, Berlin Heidelberg New York.

van Dam-Bates, P., Gansell, O. and Robertson, B. (2018). "Using balanced acceptance sampling as a master sample for environmental surveys," *Methods in Ecology and Evolution*, 9, 1718–1726.